

Distributed Community Detection in Dynamic Graphs

Andrea Clementi¹, Miriam Di Ianni¹, Giorgio Gambosi¹, Emanuele Natale², and Riccardo Silvestri³

¹ Università *Tor Vergata* di Roma, ‘‘lastname’’@mat.uniroma2.it

² Master-Degree Student, Università *Tor Vergata* di Roma, emanatale@gmail.com

³ *Sapienza* Università di Roma, silvestri@di.uniroma1.it

February 25, 2013

Abstract

The *social* behaviour of human agents in *Opportunistic Networks* has a strong impact on the performances of communication protocols since the dynamic topology of such networks is often conditioned by the presence of *Communities*, i.e., sets of agents that use to meet each other much more often than the average intermeeting time.

Community Detection is thus a preliminary task that may reveal to be crucial in order to design and analyse efficient communication protocols for opportunistic networks.

Our contribution is the first analytical study of this task in dynamic networks. We provide a framework that formalizes the distributed version of the Community-Detection Problem over a general model of dynamic networks. According to this framework, the problem turns out to be a *node-coloring* task of the dynamic graph.

Then, we present an efficient provably-good coloring protocol for two classes of dynamic random graphs that have been recently adopted as mathematical models of some opportunistic networks.

Keywords. Distributed Computing, Dynamic Graphs, Social Opportunistic Networks.

1 Introduction

Recent studies in opportunistic networks focus on the impact of the *agent social behavior* on some basic communication tasks such as routing and broadcasting [3, 13, 14]. Strong attention on this issue has been given in an emerging class of opportunistic networks called *Intermittently-Connected Mobile Networks (ICMNs)* [16]: such networks are characterized by wireless links, representing opportunities for exchanging data, that appear sporadically among humans carrying mobile radio devices.

So-called *social-aware* communication protocols rely on the reasonable intuition that since mobile devices are carried by people who tend to form *communities* then members (i.e. nodes) of the same community use to meet each other much more often than nodes from different communities. Experiments on real-data sets have widely shown that identifying communities can strongly help

improving the performance of the protocols [3, 13, 14]. It thus follows that community detection in ICMNs is a crucial issue.

Many centralized community-detection methods have been proposed in the literature (for a good survey see [7]) that may result useful for offline data analysis of mobile traces. However, it is a common belief that next-future technologies will yield a dramatic growth of *self-organizing* ICMNs where the network protocols work without relying on any centralized server. In this new communication paradigm, it is required that community detection is performed in a fully distributed way.

To the best of our knowledge, only experimental studies are available for this important task. In [10], some greedy protocols are tested on specific sets of real mobility-trace datas. By running such protocols, every node constructs and update its own community-list according to the length and the rate of the contacts observed so far by itself and by the nodes it meets. So, the protocol exploits the intuitive fact that communities are formed by nodes that use to meet often and for a long time. This assumption is somehow equivalent to the popular concept of *latent classes* introduced by Wasserman and Anderson and by Snijders and Nowicki in social-network theory (for a good survey on this topic see [9]). However, in such heuristic solutions, nodes need to update and transmit relatively large lists of node-IDs during all the process. In several real ICMNs, the overhead may result too heavy and, moreover, it may not be necessary for improving the performance of the protocols. Indeed, in several cases, it suffices to let each node to detect the community of every (dynamic) neighbor. As described in the next paragraph, this can be obtained by computing a coloring of the nodes that induces the *right* partition of the node set.

Distributed Community Detection in Dynamic Graphs. We propose a simple framework for defining the distributed Community-Detection problem in general dynamic networks.

A *dynamic graph* is a probabilistic process that describes a graph whose topology changes with time, so it can be represented by a sequence $\mathcal{G} = \{G_t = ([n], E_t) : t \in \mathbb{N}\}$ of graphs with the same set $V = [n]$ of nodes, where G_t is the *snapshot* of the dynamic graph at time step t . Thus, the presence/absence of every link changes at any time step according to a (probabilistic) process. Our framework is inspired by the “statistical” approach in [9] based on latent classes. In particular, we formalize the following two key-ideas. (i) Every snapshot of a dynamic graph can be seen as a single sample from a specific graph distribution yielded by the probabilistic process: so, any social structure of a dynamic network cannot be defined as a property of the single sample, rather it must be defined over the graph probability distribution determined by the process. (ii) Extracting information about the network structure is thus a *statistical* task which can be performed by observing a sequence of samples from the graph probability-distribution. However, differently from typical approaches in statistical inference, this task must be done in a fully-distributed way.

We can thus define the Community-Detection problem as a node-coloring problem. Let \mathcal{G} be a dynamic network whose node set $[n]$ is partitioned into V_1, \dots, V_ℓ *unknown* communities. We say that a function $Z : V \rightarrow \{1, \dots, \ell\}$ is a *good coloring* for \mathcal{G} if Z colors each community with different colors, i.e.,

$$\forall i, k \in [\ell] \forall u \in V_i \forall v \in V_k : Z(u) = Z(v) \leftrightarrow i = k$$

Then, the goal is to derive efficient distributed protocols for constructing a good coloring of a dynamic graph \mathcal{G} structured as unknown communities. Nodes are entities that share a global clock and know the number n of nodes but it is not required they have distinct IDs. We again emphasize that, initially, each node does not know its own community and it is not able to distinguish the

community of its neighbors. At every time step, every node can exchange information with all its current neighbors.

By adopting the concept of latent classes [9, 10], we first consider a simple community-network model where, at every time step, every link between two nodes of the same community is present with probability p while the link probability goes down to $q \ll p$ when the two nodes belong to different communities. This model of dynamic random network is a variant of the *dynamic Erdős–Rényi graph* model [1, 3] with a non-homogeneous edge-probability function. It will be shortly named $\mathcal{G}(n, p, q)$ where n is the number of nodes (nodes) while $p = p(n)$ and $q = q(n)$ are the edge-probability functions. This model clearly assumes important simplifications that may impact several properties of real opportunistic networks: for instance, we have assumed that contacts between nodes follow Bernoulli Processes, so the distribution of time between two contacts of a pair follows an exponential law. Previous experiments have shown that this assumption holds only at the timescale of days and weeks [3, 11]. However, in [4], experimental validations have shown that some real ICMNs (e.g. those studied in the Hagggle Project [3] and in the MIT Reality Mining Project [8]) exhibit some crucial connectivity properties (such as hop diameter) which are well-approximated by *sparse* dynamic Erdős–Rényi graphs. The model $\mathcal{G}(n, p, q)$ thus aims to remove the full-homogeneity assumption [4] of the dynamic Erdős–Rényi graph model by introducing the presence of unknown communities.

Another simplifying assumption in the dynamic Erdős–Rényi graph model is *time independence*: the graph topology at time t is fully independent from the previous topology. *Edge Markovian Evolving Graphs* (in short *edge-MEG*) were first introduced in [5] as a generalization of the dynamic Erdős–Rényi graph model that captures the strong dependence between the existence of an edge at a given time step and its existence at the previous time step. An edge-MEG is a dynamic random graph $\mathcal{G}(n, p_{\uparrow}, p_{\downarrow}, E_0) = \{G_t = ([n], E_t) : t \in \mathbb{N}\}$ defined as follows. Starting from an initial random edge set E_0 , at every time step, every edge changes its state (existing or not) according to a two-state Markovian process with probabilities p_{\uparrow} and p_{\downarrow} . If an edge exists at time t , then, at time $t + 1$, it dies with probability p_{\downarrow} . If instead the edge does not exist at time t , then it will come up at time $t + 1$ with probability p_{\uparrow} .

We observe that the setting $p_{\downarrow} = 1 - p_{\uparrow}$ yields a sequence of independent Erdős–Rényi random graphs, i.e., *dynamic Erdős–Rényi graphs* with edge probability $p = p_{\uparrow}$. Edge-MEGs have been adopted as concrete models for several real dynamic networks such as faulty networks [6], peer-to-peer systems [15], mobile ad-hoc networks [15], and vehicular networks [12]. Furthermore, Edge-MEGs have been considered by Whitbeck et al [16] as a concrete model for analyzing the performance of epidemic routing on sparse ICMNs and the obtained theoretical results have been also validated over real trace datas such as the *Rollernet* traces [14]. In this paper, we consider the Edge-MEG as a mathematical model for ICMNs. The concept of unknown communities in edge-MEGs can be introduced similarly to the way adopted for the dynamic Erdős–Rényi graphs: here, we have two edge-probability parameter pairs $(p_{\uparrow}, p_{\downarrow})$ and $(q_{\uparrow}, q_{\downarrow})$ between two nodes u and v depending on whether they both belong to the same community or not. So, if both u and v do belong to the same community then the edge (u, v) is governed by the 2-state Markov chain with parameters $(p_{\uparrow}, p_{\downarrow})$ otherwise the edge is governed by the 2-state Markov chain with parameter $(q_{\uparrow}, q_{\downarrow})$. We assume that $p_{\uparrow} \ll q_{\uparrow}$ and, according to the parameter tuning performed in [16], it turns out that the best fitting to real scenarios is achieved by setting p_{\downarrow} , (and q_{\downarrow}) as absolute constants. This is mainly due to the fact that, once a connection comes up, its expected life-time does not depend on the size of the network [16].

Our Algorithmic Contribution. We provide an efficient distributed randomized protocol for the coloring problem in the $\mathcal{G}(n, p, q)$ model with a constant number ℓ of Communities. According to the latent-class concept [9, 10] and the analysis on the clustering-degree of some real opportunistic networks performed in [17], we here assume that the edge-probability functions p and q are such that $q = O(p/n^b)$, where $b > 0$ is an *arbitrarily small* constant. Our protocol yields with high probability¹ (in short *w.h.p.*) a good coloring in $O\left(\max\{\log n, \frac{\log n}{pn}\}\right)$ time. The bound is tight for any $p = O(1/n)$ while it is only a logarithmic factor larger than the optimum for the rest of the parameter range (i.e. for more dense topologies).

The local coloring rule adopted by the protocol is simple and requires no node IDs. Furthermore, differently from the previous solutions [10], our protocol preserves the *privacy* of the nodes. Indeed, a node does not need to store or exchange any “personal” information (IDs, contact frequencies, etc) about the nodes he has met so far: the unique exchanged information is the assigned color.

Our protocol can be easily adapted in order to construct a good-coloring for the Edge-MEG model $\mathcal{G}(n, p_\uparrow, p_\downarrow, q_\uparrow, q_\downarrow, E_0)$ in the parameter range $q_\uparrow = O(p_\uparrow/n^b)$, where b is *any* positive constant. The completion time is w.h.p. bounded by

$$O\left(M \cdot \max\left\{\log n, \frac{\log n}{p_\uparrow n}\right\}\right)$$

where M is a bound on the *mixing* time of the two 2-state Markov chains governing the edges of the dynamic graph. It is known that (see for example [5])

$$M = O\left(\max\left\{\frac{1}{p_\uparrow + p_\downarrow}, \frac{1}{q_\uparrow + q_\downarrow}, \log n\right\}\right)$$

Observe that, when p_\downarrow and q_\downarrow are some arbitrary positive constants and $p_\uparrow = \Omega(1/n)$ (this case includes the “realistic” range derived in [16]), then $M = O(\log n)$ and the bound on the completion time becomes $O(\log^2 n)$. This bound is only a logarithmic factor larger than the optimal coloring time in the case of sparse topologies, i.e., when $p_\uparrow = \Theta(1/n)$.

We run our protocol over hundreds of random instances according to the $\mathcal{G}(n, p, q)$ model with n varying from 10^3 to 10^6 . Besides a good validation of our asymptotical analysis, further positive features of the protocol are shown. Our protocol is indeed *tolerant* to non-homogeneous edge-probability functions. In particular, the protocol almost-always returns a good coloring in *Bernoullian* graphs where the edge probability is not uniform, i.e., for each pair (u, v) of nodes in the same community, the parameter $p_{u,v}$ is suitably chosen in order to yield irregular sparse graphs. A detailed description of the experimental results can be found in the Appendix (Section A).

2 The protocol and its analysis

In this section, we first consider the dynamic graph $\mathcal{G}(n, p, q)$ and, for the sake of clarity, we assume the following restrictions hold: the parameter p is known by every node; there are only 2 communities V_1 and V_2 , each of size $n/2$ (n is an even number); the coloring process starts with (exactly) two *source nodes*, $s_1 \in V_1$ that is z_1 -colored and $s_2 \in V_2$ that is z_2 -colored with $z_1 \neq z_2$. The parameters p and q belong to the following ranges

$$\frac{1}{n} \leq p \leq \frac{d \log n}{n}, \text{ for some constant } d > 0 \text{ and } q = O\left(\frac{p}{n^b}\right), \text{ for some constant } b > 0 \quad (1)$$

¹As usual, we say an event holds with high probability if it holds with probability not smaller than $1 - \frac{1}{n^{\Theta(1)}}$.

Such restrictions make the description much easier, thus allowing us to focus on the main ideas of our protocol and of its analysis. Then, at the end of this section, we will show how to remove the assumption on the presence of the two sources and, in Section 3, we will get the general result stated in the introduction.

The protocol relies on the following simple and natural idea. Starting from two source nodes (one in each community), each one having a different color, the protocol performs a color spreading by adopting a simple coloring/broadcasting rule (for instance, every node gets the color it sees most frequently in its neighbors). Since links between agents of the same community are much more frequent than the others, we can argue that the *good-coloring* will be faster than the *bad-coloring* (in each community, the good coloring is the one from the source of the community while the bad coloring is the one coming from the other source). However, providing a rigorous analysis of the above process requires to cope with some non-trivial probabilistic issues that have not been considered in the analysis of information spreading in dynamic graphs made in previous works [2, 5, 6]. Let us consider any local coloring rule that (only) depends on the color configuration of the (dynamic) neighborhood of the node. At a given time step, there is a subset $I_c \subseteq [n]$ of colored nodes and we need to evaluate the probabilities P_g (P_b) that a non-colored node gets a good (bad) color in the next step. After an initial phase, there is a non-negligible probability that some nodes will get the bad color. Then, such nodes will start a spreading of the bad coloring at the same rate of the good one. It turns out that the probabilities P_g and P_b strongly depend on the *balance* between the sizes of the subsets of well-colored nodes (and the badly-colored ones) reached in the two communities, respectively. Keeping a tight balance between such values during all the process is the main technical goal of the protocol. In arbitrary color configurations over *sparse* graph snapshots, getting “high-probability” bounds on the rate of new (well/badly) colored nodes is a non-trivial issue. Moreover, it is not hard to show that, given any two nodes $v, w \in [n] \setminus I_c$, the events “ v will be (well/badly-)colored” and “ w will be (well/badly-)colored” are not independent. As we will see, such issues are already present in the “restricted” case considered in this section. A first important step of our approach is to describe the combination between the coloring process and the dynamic graph as a finite-state Markovian process. Then, we perform a step-by-step analysis, focusing on the probability that the Markovian Process visits a sequence of states having “good-balance” properties.

Our protocol applies local rules only depending on the current node’s neighborhood and on the current time step. The protocol execution over the dynamic graph can be represented by the following *Markovian Process*: for any time step t , we denote as $\left(k_1^{(t)}, k_2^{(t)}, h_1^{(t)}, h_2^{(t)}; E_t\right)$ the *state* reached by the Markovian Process where $k_i^{(t)}$ denotes the number of nodes in the i -th community colored by color z_i at time step t and $h_i^{(t)}$ denotes the number of nodes in the i -th community colored by color z_j at time step t , for $i, j = 1, 2$ and $j \neq i$. In particular, the Markovian Process works as follows

$$\begin{aligned} \dots \rightarrow \left(k_1^{(t)}, k_2^{(t)}, h_1^{(t)}, h_2^{(t)}; E_t\right) &\xrightarrow{\mathcal{G}(n, p, q)} \left(k_1^{(t)}, k_2^{(t)}, h_1^{(t)}, h_2^{(t)}; E_{t+1}\right) \xrightarrow{\text{protocol}} \\ &\xrightarrow{\text{protocol}} \left(k_1^{(t+1)}, k_2^{(t+1)}, h_1^{(t+1)}, h_2^{(t+1)}; E_{t+1}\right) \xrightarrow{\mathcal{G}(n, p, q)} \dots \end{aligned}$$

The major profit of this description is that, if we observe the process in any fixed state, then it is not hard to verify that, in the next time step, the events {“node v gets a good/bad color”,

$v \in V$ }, become mutually independent. This will allow us to preserve good-balance properties for a sufficiently-long sequence of states visited by the Markovian Process, w.h.p.

The protocol works in 5 consecutive temporal phases: the goal of this phase partition is to control the rate of new colored nodes as function of the expected values reached by the r.v. $k_i^{(t)}, h_i^{(t)}$ (at the end of each phase). Indeed, when such expected values reach some specific thresholds, the protocol and/or its analysis must change accordingly in order to keep the coloring configuration well-balanced in the two communities during all the process.

At any time step t , we denote, for each node $v \in V_1$, the number of z_1 -colored neighbors of v as $N_1^v(t)$; $N_2^v(t)$ is defined similarly as the number of z_2 -colored neighbors of v . Given a node $v \in V$, the set of its neighbors at time t will be denoted as $\Gamma_t(v)$. For the sake of brevity, whenever possible we will omit the parameter t in the above variables and, in the proofs, we will only analyze the coloring in V_1 , the analysis for V_2 being the same.

2.1 Phase 1: source coloring

The phase runs for $\tau_1 = c_1 \log n$ time steps, where $c_1 > 0$ is an explicit constant that will be fixed later. In this phase, only the neighbors of the sources will decide their color. The goal is to reach a state such that w.h.p. $k_i = \Theta(\log n)$ and $h_i = 0$ ($i = 1, 2$). For any non-source node v , the coloring rule is thus the following.

- Let $i \in \{1, 2\}$; v gets color z_i if there is a time step $t \leq \tau_1$ such that $s_i \in \Gamma_t(v)$ and, for $j \neq i$ and for all t such that $1 \leq t \leq \tau_1$, it holds that $s_j \notin \Gamma_t(v)$;
- In all other cases, v gets no color.

Thus, in order to analyze the process, it is appropriate to define the following r.v.s counting the colored nodes at the end of Phase 1.

- The variable $X_1^v = 1$ iff v gets color z_1 , and the variable $X_1 = 1 + \sum_{v \neq s_1} X_1^v$ describes the total number of the z_1 -colored nodes in V_1 .
- The variable $Y_1^v = 1$ iff v gets color z_2 and the variable $Y_1 = \sum_{v \neq s_1} Y_1^v$ describes the total number of the z_2 -colored nodes in V_1 .

The key property of the protocol in this phase is stated in the following theorem.

Theorem 1 *Let $d_1 > 0$ be any (sufficiently large) constant. Then, a constant $c_1 > 0$ can be fixed so that, at time step $\tau_1 = c_1 \log n$ the Markovian Process w.h.p. reaches a state such that*

$$k_1^{(\tau_1)}, k_2^{(\tau_1)} \in \left[\frac{d_1}{16} pn \log n, 4d_1 pn \log n \right] \quad \text{and} \quad h_1^{(\tau_1)}, h_2^{(\tau_1)} = 0 \quad (2)$$

Proof. The above theorem follows from the fact that, at the end of this phase, a node gets the good color with probability $\Theta(p\tau_1)$ and, w.h.p., no node will get the bad color.

Claim 1 Let τ_1 be such that $p\tau_1 \geq \frac{\log n}{n}$ and $\tau_1 = o\left(\frac{1}{p}\right)$. Then, starting from the initial state $(k_1^{(0)} = 1, k_2^{(0)} = 1, h_1^{(0)} = 0, h_2^{(0)} = 0; E_0)$, at time step τ_1 w.h.p. it holds that

$$\frac{1}{16}np\tau_1 \leq X_1, X_2 \leq 4np\tau_1$$

Proof. We first bound from below the number of z_1 -colored nodes at the end of Phase 1. Notice that $\mathbf{P}(X_1^v = 1) = (1 - q)^{\tau_1} (1 - (1 - p)^{\tau_1})$ and we can apply Lemma 10 to each factor on the right side, thus getting:

$$\mathbf{P}(X_1^v = 1) = (1 - q)^{\tau_1} (1 - (1 - p)^{\tau_1}) \geq p\tau_1 [(1 - q\tau_1(1 + 2q))(1 - p\tau_1)] \geq \frac{p\tau_1}{2}$$

where in the last inequality we used $\lim_{n \rightarrow \infty} [(1 - q\tau_1(1 + 2q))(1 - p\tau_1)] = 1$.

The above inequality easily implies that

$$\mathbf{E}[X_1] = 1 + (|V_1| - 1) \cdot \mathbf{P}(X_1^v = 1) \geq 1 + \left(\frac{n}{2} - 1\right) \frac{p\tau_1}{2} > \frac{n}{2} \left(1 - \frac{2}{n}\right) \frac{p\tau_1}{2},$$

that is

$$\mathbf{E}[X_1] > \frac{np\tau_1}{8}$$

Since, by fixing any initial state $(k_1^{(0)}, k_2^{(0)}, h_1^{(0)}, h_2^{(0)}; E_0)$, the random variables X_1^v are independent, we can apply the Chernoff Bound (17) with $\delta = \frac{1}{2}$. Then,

$$\mathbf{P}\left(X_1 \leq \frac{np\tau_1}{16}\right) \leq e^{-\frac{1}{64}np\tau_1}$$

By hypothesis, we have that $np\tau_1 \geq \log n$, so, w.h.p. it holds that

$$X_1 \geq \frac{np\tau_1}{16}$$

A similar analysis, based on Lemma 10 and Chernoff bound (18), yields the stated upper bound on the number of z_1 -colored nodes at the end of the Phase 1, that is, w.h.p. it holds that

$$X_1 \leq 4np\tau_1$$

□

Claim 2 Let $\tau_1 \geq 1$ be such that $q\tau_1 = O\left(\frac{1}{n^{1+\epsilon}}\right)$ for any $\epsilon > 0$. Then, starting from the initial state $(k_1^{(0)} = 1, k_2^{(0)} = 1, h_1^{(0)} = 0, h_2^{(0)} = 0; E_0)$, at time step τ_1 it holds w.h.p. that $Y_1 = 0$.

Proof. A sufficient condition for the event $Y_1 = 0$ is that no edge between any node in V_1 and s_2 occurs at any time step of Phase 1. Hence, by Lemma 10,

$$\mathbf{P}(Y_1 = 0) \geq (1 - q)^{|V_1|\tau_1} = (1 - q)^{\frac{n\tau_1}{2}} \geq 1 - 2q|V_1|\tau_1$$

Since $q\tau_1 = O\left(\frac{1}{n^{1+b}}\right)$, the lemma is proved.

□

Claims 1 and 2 easily imply Theorem 1. □

2.2 Phase 2: fast coloring I

This phase of the Protocol aims to get an exponential rate of the good-coloring inside every community in order to reach, in $\tau_2 = O(\log n)$ time, a state such that the number of well-colored nodes is bounded by some root of n and the number of badly-colored ones is still 0.

We consider the Markovian Process when, at the generic step t of this phase, it is in any state satisfying the following condition:

$$k_1^{(t)}, k_2^{(t)} \in \left[\frac{d_1}{16} pn \log n, d_1 pn^{1+a} \log n \right] \text{ and } h_1^{(t)}, h_2^{(t)} = 0 \quad (3)$$

Differently from Phase 1, nodes can get a color at every time step according to the following rule: for $\tau_1 < t \leq \tau_1 + \tau_2$, at time step t of Phase 2 every *uncolored* node v

- gets color z_1 at time $t + 1$ iff $N_1^v(t) > 0$ and $N_2^v(t) = 0$,
- gets color z_2 at time $t + 1$ iff $N_2^v(t) > 0$ and $N_1^v(t) = 0$,
- gets no color at time $t + 1$ otherwise.

For each time step t , $\tau_1 < t \leq \tau_1 + \tau_2$, we thus define the following binary random variables

- $X_1^v(t) = 1$ iff $v \in V_1$ gets color z_1 at time $t + 1$, and $X_1(t) = \sum_{v \in V_1} X_1^v(t)$.
- $Y_1^v(t) = 1$ iff $v \in V_1$ gets color z_2 at time $t + 1$, and $Y_1(t) = \sum_{v \in V_1} Y_1^v(t)$

In the next theorem, we assume that, at time step τ_1 (i.e. at the end of Phase 1), the Markovian Process reaches a state satisfying Cond. (2). In particular, we assume that $k_i^{\tau_1} \geq \underline{k}_i^{\tau_1}$, where $\underline{k}_i^{\tau_1} = \frac{d_1}{16} pn \log n$. Thanks to Theorem 1, this event holds w.h.p. In what follows, we will make use of the following function

$$F(n, k) = 2 \max \left\{ \sqrt{\frac{\log n}{k}}, \frac{\text{polylog } n}{n^{1-a}} \right\}$$

Theorem 2 *For any $\eta > 0$, positive constants a and ϕ can be fixed so that, at the final step of Phase 2*

$$\tau_2 = \frac{1}{\log \left(1 + \left(\frac{np}{2} \right) \right)} \log \left(\frac{n^a}{\phi \log^3 n} \right) + \log^{-1} \left[\left(1 + \frac{np}{2} \right) \left(1 - F(n, \underline{k}_1^{(\tau_1)}) \right) \right] \log \left(\frac{\log^3 n}{\underline{k}_1^{(\tau_1)}} \right) + \tau_1,$$

it holds w.h.p. that

$$\text{for } i = 1, 2, \quad n^a \leq k_i^{(\tau_2)} \leq n^a \log^\eta n, \quad \text{and } h_i^{(\tau_2)} = 0. \quad (4)$$

Proof. We prove the following key-fact: if $k_i = O(n^a)$ ($i = 1, 2$) for some constant $a < 1$, then it holds w.h.p. $X = [(1 \pm o(1))pn/2] \cdot k_1$ and $Y = 0$. From such bounds, we then derive the recursive equations for $k_i^{(t)}$ yielding the bounds stated in the theorem.

Remind that $q = O(p/n^b)$ and consider any positive constant a such that $a < b$.

We first provides tight upper and lower bounds on the number of new colored nodes after one step of the protocol.

Claim 3 *For $i = 1, 2$, it holds w.h.p. that*

$$\left(1 - \sqrt{\frac{\log n}{k_i^{(t)}}}\right) \left(1 - \frac{\text{polylog } n}{n^{1-a}}\right) \frac{np}{2} k_i^{(t)} \leq X_i(t) \leq \left(1 + \sqrt{\frac{\log n}{k_i^{(t)}}}\right) \left(1 + \frac{\text{polylog } n}{n^{1-a}}\right) \frac{np}{2} k_i^{(t)}$$

$$Y_i(t) = 0.$$

Proof. Observe that $\mathbf{P}(Y_1(t) = 0)$ is lower bounded by the probability that in E_t there is no edge between any node in V_1 and any node in V_2 which is already colored z_2 . By the hypothesis (3) and the conditions on p and q , we can thus apply Lemma 10 and get

$$\mathbf{P}(Y_1 = 0) \geq (1 - q)^{|V_1|k_2^{(t)}} \geq 1 - 2q|V_1|k_2^{(t)} \geq 1 - \frac{\text{polylog } n}{n^{1-a}}$$

proving that w.h.p. $Y_1(t) = 0$. Again, thanks to Condition (3) and the conditions on p and q (Eq. 1), we can apply Lemma 10 to bound $\mathbf{P}(X_1^v = 1) = \left(1 - (1 - p)^{k_1^{(t)}}\right) (1 - q)^{k_2^{(t)}}$.

We get

$$\mathbf{P}(X_1^v = 1) \geq k_1^{(t)} p \left(1 - k_1^{(t)} p\right) \left(1 - 2k_2^{(t)} q\right) \geq k_1^{(t)} p \left(1 - \frac{\text{polylog } n}{n^{1-a}}\right)$$

$$\mathbf{P}(X_1^v = 1) \leq k_1^{(t)} p (1 + 2p) \leq k_1^{(t)} p \left(1 + \frac{\text{polylog } n}{n^{1-a}}\right)$$

We can thus bound the expected number of new well-colored nodes $\mathbf{E}[X_1] = (|V_1| - k_1^{(t)}) \mathbf{P}(X_1^v = 1)$:

$$\frac{np}{2} k_1^{(t)} \left(1 - \frac{\text{polylog } n}{n^{1-a}}\right) \leq \mathbf{E}[X_1] \leq \frac{np}{2} k_1^{(t)} \left(1 + \frac{\text{polylog } n}{n^{1-a}}\right)$$

By applying the Chernoff Bounds (17) and ((18)) with $\delta = \sqrt{\frac{\log n}{k_1}}$, we get

$$\mathbf{P}\left(X_1 \leq \left(1 - \sqrt{\frac{\log n}{k_1^{(t)}}}\right) \left(1 - \frac{\text{polylog } n}{n^{1-a}}\right) \frac{np}{2} k_1^{(t)}\right) = e^{-\frac{\log n}{2k_1^{(t)}} \frac{np}{2} k_1^{(t)} \left(1 - \frac{\text{polylog } n}{n^{1-a}}\right)} \leq \frac{1}{n^{\frac{1}{3}}}$$

$$\mathbf{P}\left(X_1 \geq \left(1 + \sqrt{\frac{\log n}{k_1^{(t)}}}\right) \left(1 + \frac{\text{polylog } n}{n^{1-a}}\right) \frac{np}{2} k_1^{(t)}\right) = e^{-\frac{\log n}{3k_1^{(t)}} \frac{np}{2} k_1^{(t)} \left(1 - \frac{\text{polylog } n}{n^{1-a}}\right)} \leq \frac{1}{n^{\frac{1}{3}}}$$

This implies that, w.h.p.,

$$\left(1 - \sqrt{\frac{\log n}{k_1^{(t)}}}\right) \left(1 - \frac{\text{polylog } n}{n^{1-a}}\right) \frac{np}{2} k_1^{(t)} \leq X_1(t) \leq \left(1 + \sqrt{\frac{\log n}{k_1^{(t)}}}\right) \left(1 + \frac{\text{polylog } n}{n^{1-a}}\right) \frac{np}{2} k_1^{(t)}$$

□

Let us observe that $k_i^{(t+1)} = k_i^{(t)} + X_i(t)$. So, Claim 3 easily implies the following recursive bounds

Claim 4 *For $i = 1, 2$, it holds w.h.p. that*

$$\left(1 + \frac{np}{2}\right) \left[1 - F(n, k_i^{(t)})\right] k_i^{(t)} \leq k_i^{(t+1)} \leq \left(1 + \frac{np}{2}\right) \left[1 + F(n, k_i^{(t)})\right] k_i^{(t)}$$

$$\text{and } h_i^{(t+1)} = 0.$$

We can now analyze the spreading of the good coloring. The idea is to derive the closed formula corresponding to the recurrence relation provided by Claim 4 and analyze it in two different spans of time: in the first span, we let k_1 increase enough so that, in the second span, we can apply a stronger concentration result. Recall that, thanks to Theorem 1, Phase 2 starts with the Markovian Process in a state satisfying Condition (2) that implies Condition (3). Moreover, we will fix the final time step τ_2 so that Condition (3) has been holding for *all time steps* of Phase 2: this implies that we can apply Claim 4 for all such steps.

Let $t^* \leq \tau_1$ be defined as

$$t^* = \log^{-1} \left(\left(1 + \frac{np}{2}\right) \left(1 - F(n, k_1^{(\tau_1)})\right) \right) \log \left(\frac{\log^3 n}{k_1^{(\tau_1)}} \right) + \tau_1.$$

Since $t^* - \tau_1 \in O(\log n)$, thanks to Lemma 12, we can unroll (backward) the recursive relation from time t^* to time τ_1 and get

$$\left(1 + \frac{np}{2}\right)^{t^* - \tau_1} \left[1 - F(n, k_1^{(\tau_1)})\right]^{t^* - \tau_1} k_1^{(\tau_1)} \leq k_1^{(t^*)} \leq \left(1 + \frac{np}{2}\right)^{t^* - \tau_1} \left[1 + F(n, k_1^{(\tau_1)})\right]^{t^* - \tau_1} k_1^{(\tau_1)} \quad (5)$$

We observe that the value of $k_1^{\tau_1} \geq \frac{d_1}{16} pn \log n$ can reach any arbitrarily large constant by tuning the constant d_1 in Theorem 1; so, $F(n, k_1^{(\tau_1)})$ can be made arbitrarily small. From this fact and Eq. 5, we have that $k_1^{(t^*)} \in [\log^3 n, \log^{3+\mu} n]$, where μ can be made arbitrarily small by decreasing $F(n, k_1^{(\tau_1)})$ (i.e. by increasing d_1 in Theorem 1). Notice that, at any time step $t \leq t^*$, Condition (3) is largely satisfied.

We now unroll the recursive relation from time τ_2 to time t^* and get

$$\left(1 + \frac{np}{2}\right)^{\tau_2 - t^*} \left[1 - F(n, k_1^{(t^*)})\right]^{\tau_2 - t^*} k_1^{(t^*)} \leq k_1^{(\tau_2)} \leq \left(1 + \frac{np}{2}\right)^{\tau_2 - t^*} \left[1 + F(n, k_1^{(t^*)})\right]^{\tau_2 - t^*} k_1^{(t^*)}. \quad (6)$$

We observe that, with a suitable choice of the positive constant $\phi \in (0, 1)$, for

$$\tau_2 = \log^{-1} \left(1 + \frac{np}{2}\right) \log \left(\frac{n^a}{\phi \log^3 n}\right) + t^*$$

it holds that

$$\left[1 - F(n, k_1^{(t^*)})\right]^{\tau_2 - t^*} \geq \phi \quad \text{and} \quad \left[1 + F(n, k_1^{(t^*)})\right]^{\tau_2 - t^*} \leq \frac{1}{\phi}$$

By replacing τ_2 into Eq. 6, with a suitable choice of $\eta \geq \mu$ (remind that μ can in turn be made arbitrarily small), we finally get $n^a \leq k_1^{(\tau_2)} \leq n^a \log^\eta n$. Again, observe that, for all time steps $t \leq \tau_2$, Condition (3) is largely satisfied: this implies that at each of these steps we were able to apply Claim 4.

As for the bad coloring, observe that Claim 4 guarantees (w.h.p.) $h_1^{(t)}, h_2^{(t)} = 0$; then, from Lemma 12, it holds w.h.p that $h_1^{(\tau_2)} = 0$ and $h_2^{(\tau_2)} = 0$. □

2.3 Phase 3: fast coloring II

In this phase nodes apply the same rule of Phase 2 but we need to separate the analysis from the previous one since, when the “well-colored” subset gets size larger than some root of n , we cannot anymore exploit the fact that the bad coloring is w.h.p. not started yet (i.e. $h = 0$). However, we will show that when the well-colored sets get size $\Theta(n/\text{polylog } n)$, the bad-colored sets have still size bounded by some root of n . We assume that, at the end of Phase 2, the Markovian Process reaches a state satysfying Cond. (4) of Theorem 2 and that , at the generic step t of current phase, it is in any state satisfying the following condition

$$\text{for } i = 1, 2 : k_i^{(t)} \in \left[n^a, \frac{n}{\log^2 n}\right] \quad \text{and} \quad h_i^{(t)} = O(n^{a_2}), \text{ where } a_1 < a_2 < 1 \quad (7)$$

Theorem 3 *For any constant $\eta > 0$, constants $a_1 < 1$ and $\gamma > 0$ can be fixed so that at the final time step of Phase 3*

$$\tau_3 = \frac{1}{\log\left(1 + \left(\frac{n^p}{2}\right)\right)} \log\left(\frac{n^{1-a}}{\gamma \log^3 n}\right) + \tau_2$$

for $i = 1, 2$, it holds w.h.p. that

$$\frac{n}{\log^3 n} \leq k_i^{(\tau_3)} \leq \frac{n}{\log^{3-\eta} n} \quad \text{and} \quad h_i^{(\tau_3)} \leq n^{a_1} \quad (8)$$

Proof. Let X and Y be the r.v.s defined in the previous subsection. The presence of the bad coloring changes the bounds we obtain as follows. At time step $t + 1$, as long as $k_i, h_i = O(n/\text{polylog } n)$, we prove that $X = [(1 \pm o(1))pn/2] \cdot k_1$ and $Y = [(1 \pm o(1))(ph_1 + qk_2)]n/2$. From the above bounds, we then determine two time-recursive bounds on the r.v. k_i^t and h_i^t that hold (w.h.p.) for any t s.t. $k_i^t, h_i^t = O(n/\text{polylog } n)$. Then, thanks to the hypothesis $q = O(p/n^b)$ and to the fact that the Markovian Process starts Phase 3 from a very “unbalanced” state ($k_i = \Omega(n^a)$ and $h_i = 0$), we apply the recursive bounds and show that a time step τ_3 exists satysfying Eq. 8.

We start by providing tight upper and lower bounds on the number of the well-colored nodes at a generic step of Phase 3 (some of the proofs are similar to those of Phase 2, so they will only sketched).

Claim 5 A constant $\zeta > 0$ exists such that, for $i = 1, 2$, it holds w.h.p. that

$$\left(1 + \frac{np}{2}\right) \left(1 - \frac{\zeta}{\log n}\right) k_i^{(t)} \leq k_i^{(t+1)} \leq \left(1 + \frac{np}{2}\right) \left(1 + \frac{\zeta}{\log n}\right) k_i^{(t)} \quad (9)$$

Sketch of Proof. By neglecting the contribution of h_2 , from the facts $pk_1^{(t)}, qk_2^{(t)}, ph_1^{(t)} = o(1)$ and Lemma 10, we have that

$$\begin{aligned} \mathbf{P}(X_1^v = 1) &\geq \left(1 - (1-p)^{k_1^{(t)}}\right) (1-q)^{k_2^{(t)}} (1-p)^{h_1^{(t)}} \\ &\geq pk_1^{(t)} \left(1 - pk_1^{(t)}\right) \left(1 - 2qk_2^{(t)}\right) \left(1 - 2ph_1^{(t)}\right). \end{aligned}$$

Observe that

$$\left(1 - pk_1^{(t)}\right) \left(1 - 2qk_2^{(t)}\right) \left(1 - 2ph_1^{(t)}\right) \geq \left(1 - \Theta\left(\frac{1}{\log^2 n}\right)\right)$$

then

$$\mathbf{P}(X_1^v = 1) \geq pk_1^{(t)} \left(1 - \Theta\left(\frac{1}{\log^2 n}\right)\right) \quad (10)$$

We now provide an upper bound on $\mathbf{P}(X_1^v = 1)$. From the Union Bound and Lemma 10, we get

$$\mathbf{P}(X_1^v = 1) \leq pk_1^{(t)} (1 + 2p) + qh_2^{(t)} (1 + 2q) \leq pk_1^{(t)} \left(1 + \Theta\left(\frac{1}{n^{1-a_2}}\right)\right). \quad (11)$$

As usual, we exploit Eq.s 10 and 11 to bound the expectation

$$\mathbf{E}[X_1] = \left(|V_1| - (k_1^{(t)} + h_1^{(t)})\right) \cdot \mathbf{P}(X_1^v = 1)$$

For some constant $\tilde{\zeta} > 0$, w.h.p. it thus holds that

$$\frac{np}{2} k_1^{(t)} \left(1 - \frac{\tilde{\zeta}}{\log n}\right) \leq \mathbf{E}[X_1] \leq \frac{np}{2} k_1^{(t)} \left(1 + \frac{\tilde{\zeta}}{\log n}\right)$$

We can use the Chernoff Bounds (17 and 18 with $\delta = 1/\log n$), to get that, for some constant $\zeta > 0$, w.h.p.

$$\frac{np}{2} k_1^{(t)} \left(1 - \frac{\zeta}{\log n}\right) \leq X_1 \leq \frac{np}{2} k_1^{(t)} \left(1 + \frac{\zeta}{\log n}\right) \quad (12)$$

From the above inequality, it follows that

$$\left(1 + \frac{np}{2}\right) \left(1 - \frac{np}{2 + np} \frac{\zeta}{\log n}\right) k_1^{(t)} \leq k_1^{(t+1)} \leq \left(1 + \frac{np}{2}\right) \left(1 + \frac{np}{2 + np} \frac{\zeta}{\log n}\right) k_1^{(t)},$$

Since $\frac{np}{2+np}\zeta$ is bounded by a constant, for the sake of simplicity we can just re-define ζ as any fixed constant such that

$$\left(1 + \frac{np}{2}\right) \left(1 - \frac{\zeta}{\log n}\right) k_1^t \leq k_1^{t+1} \leq \left(1 + \frac{np}{2}\right) \left(1 + \frac{\zeta}{\log n}\right) k_1^t$$

□

Claim 5 implies the following properties of the well-coloring for any state within Phase 3 (including the final one at time τ_3).

Claim 6 *A constant $\zeta > 0$ exists such that, for $i = 1, 2$, it holds w.h.p. that*

$$\left(1 + \frac{np}{2}\right)^{t+1-\tau_2} \left(1 - \frac{\zeta}{\log n}\right)^{t+1-\tau_2} k_1^{(\tau_2)} \leq k_1^{(t+1)} \leq \left(1 + \frac{np}{2}\right)^{t+1-\tau_2} \left(1 + \frac{\zeta}{\log n}\right)^{t+1-\tau_2} k_1^{(\tau_2)}$$

Sketch of Proof. Since Claim 5 holds as long as $k_i \leq \frac{n}{\log^2 n}$, from Lemma 12 we get the claim by applying the same unrollement argument shown in the previous phase. □

We now exploit the above lemma to provide a bound on the number of bad-colored nodes at the end of Phase 3.

Claim 7 *For any positive constant γ , a constant a_1 , with $1 - a < a_1 < a_2$, can be fixed so that by choosing the final time step of Phase 3*

$$\tau_3 = \frac{1}{\log\left(1 + \left(\frac{np}{2}\right)\right)} \log\left(\frac{n^{1-a}}{\gamma \log^3 n}\right) + \tau_2,$$

it holds w.h.p. that, for $i = 1, 2$ and for all $t \leq \tau_3$, $h_i^{(t)} \leq n^{a_1}$.

Sketch of Proof. In order to bound the rate of $h_1^{(t)}$, we consider the r.v. Y_1^v when the Markovian Process is in a generic state satisfying Condition (7). Thanks to Theorem 2, we know this (largely) holds for the first step of Phase 3 and, by the choice of τ_3 , we will see this (w.h.p.) holds for all $t \leq \tau_3$ by induction.

By neglecting the contribution of k_2 , we have that

$$\mathbf{P}(Y_1^v = 1) \geq \left(1 - (1-p)^{h_1^{(t)}}\right) (1-q)^{h_2^{(t)}} (1-p)^{k_1^{(t)}}$$

Since $ph_1^{(t)}, qh_2^{(t)}, pk_1^{(t)} = o(1)$, we can apply Lemma 10 to each factor on the right side, thus obtaining

$$\begin{aligned} \mathbf{P}(Y_1^v = 1) &\geq ph_1^t \left(1 - ph_1^{(t)}\right) \left(1 - 2qh_2^{(t)}\right) \left(1 - 2pk_1^{(t)}\right) \\ &\geq ph_1^t \left(1 - \Theta\left(\frac{1}{\log^2 n}\right)\right) \end{aligned}$$

We now provide an upper bound to $\mathbf{P}(Y_1^v = 1)$. By the Union Bound and Lemma 10 we get

$$\begin{aligned}\mathbf{P}(Y_1^v = 1) &\leq \left(1 - (1-p)^{h_1^{(t)}}\right) + \left(1 - (1-q)^{k_2^{(t)}}\right) \\ &\leq ph_1^{(t)}(1+2p) + qk_2^{(t)}(1+2q) \\ &\leq (ph_1^{(t)} + qk_2^{(t)})(1+2p)\end{aligned}$$

As for the expected value of new bad-colored nodes, for some constant $\zeta > 0$, it holds that

$$\frac{np}{2}h_1^{(t)}\left(1 - \frac{\zeta}{\log n}\right) \leq \mathbf{E}[Y_1] \leq \left(\frac{np}{2}h_1^{(t)} + \frac{nq}{2}k_2^{(t)}\right)\left(1 + \frac{1}{\log n}\right) \quad (13)$$

From the Chernoff Bound and Eq. (13), it follows that w.h.p. h_1 will not “jump” from a sublogarithmic value to a polynomial one: in other words, in the first time that T will be at least $\log^3 n$, we have that $h_1^T = O(\text{polylog } n)$.

Hence, again from the Chernoff Bound and Eq. (13), setting $\delta = \sqrt{\frac{\log n}{h_1^{(t)}}}$, we see that for each $t \geq T$ in Phase 3 w.h.p. we have

$$Y_1 \leq \left(\frac{np}{2}h_1^{(t)} + \frac{k_2^{(t)}}{2n^\alpha}\right)\left(1 + \frac{2}{\log n}\right) \quad (14)$$

In Eq. (14), we can bound the term $\frac{k_2^{(t)}}{2n^\alpha}$ using Claim 6 and Theorem 2, obtaining for some positive constant c

$$\frac{k_2^{(t)}}{2n^\alpha} \leq \left(1 + \frac{np}{2}\right)^{t-\tau_2} \frac{\left(1 + \frac{\zeta}{\log n}\right)^{t-\tau_2} k_1^{(\tau_2)}}{2n^\alpha} \leq \left(1 + \frac{np}{2}\right)^{t-\tau_2} \cdot c$$

Therefore we can use Eq. (14) to get that w.h.p.

$$h_1^{(t+1)} = h_1^t + Y_1 \leq \left(\left(1 + \frac{np}{2}\right)h_1^{(t)} + \left(1 + \frac{np}{2}\right)^{t-\tau_2} \cdot c\right)\left(1 + \frac{2}{\log n}\right) \quad (15)$$

Hence unrolling $h_1^{(t)}$ until time T , for some positive constants c_1 and c_2 , Eq. (15) becomes (keeping high probability thanks to Lemma 12)

$$\begin{aligned}h_1^{t+1} &\leq \left(1 + \frac{np}{2}\right)^{t+1-T} \cdot \left(1 + \frac{2}{\log n}\right)^{t+1-T} \cdot h_1^{(T)} + c \cdot \left(1 + \frac{np}{2}\right)^{t-\tau_2} \cdot \sum_{i=T}^t \left(1 + \frac{2}{\log n}\right)^{t+1-i} \\ &\leq c_1 \left(1 + \frac{np}{2}\right)^{t+1-T} h_1^{(T)} + c_2 \log n \left(1 + \frac{np}{2}\right)^{t-\tau_2}\end{aligned}$$

and the last side turns out to be $O(n^{1-a} \cdot \text{polylog } n)$ when $t+1 = \tau_3$, proving the lemma. \square

The bound claimed for h_i follows from Claim 7, and the bounds claimed for k_i follow from Claim 6 for $t = \tau_3$, thus proving Theorem 3 \square

Theorems 2 and 3 guarantee a very tight range for the r.v. k_1 and k_2 at the final step of Phase 2 and 3, respectively. As we will see later, this tight balance is crucial for removing the hypothesis on the existence of the two leaders.

2.4 Phase 4: controlled saturation

At the end of Phase 3, the Markovian Process w.h.p. reaches a state that satisfies the properties stated in Theorem 3. The goal of Phase 4 is to obtain a (large) constant fraction α (say, $\alpha = 3/4$) of the nodes of each community that get the good color and, at the same time, to ensure that the number of bad-colored nodes is still bounded by some root of n . We cannot guarantee this goal by applying the same coloring rule of the previous phase: the number of bad-colored nodes would increase too fast. The protocol thus performs a much “weaker” coloring rule that is enough for the good coloring but it keeps the final number of bad-colored nodes bounded by some root of n .

The fourth phase consists of three consecutive identical time-windows during which *every* (colored or not) node $v \in V$ applies the following simple rule:

For any $t \in [\tau_3 + 1, \tau_3 + T_4 = c_4 \log n]$, v looks at the colors of its neighbors at time t and:

- If v sees only one color (say, z) for *all* the window’s time steps, then v gets color z ;
- In all the other cases (either v sees more colors or v does not see any color), v keeps its color (if any) or it remains uncolored.

The above Protocol window is repeated 3 times for a specific setting of the constant c_4 that will be determined in the proof of Theorem 4. We assume that the Markovian Process terminates Phase 3 reaching a state that satisfies Eq. 8: thanks to Theorem 3, we know this holds w.h.p.

Theorem 4 *Let α be any constant such that $0 < \alpha < 1$. Then, constants c_4 and $a_1 < 1$ can be fixed so that, at time step $\tau_4 = \tau_3 + 3T_4$, the Markovian Process w.h.p reaches a state such that, for $i = 1, 2$,*

$$k_i^{\tau_4} \geq \alpha n, \text{ and } h_i^{\tau_4} \leq n^{a_1} \text{ polylog } n. \quad (16)$$

Proof.

We first bound the number of new bad-colored nodes.

Claim 8 *For any constant c_4 , at time step $\tau_4 = 3T_4 + \tau_3 = 3c_4 \log n + \tau_3$, the Markovian Process is w.h.p. in a state such that $h_i^{\tau_4} = O(n^{a_1} \text{ polylog } n)$, for $i = 1, 2$.*

Sketch of Proof. Let $h_1 = h_1^{(\tau_3)}$ and $k_2 = k_2^{(\tau_3)}$. At the end of the first window, for any node $v \in V_1$, it holds that

$$\begin{aligned} \mathbf{P}(Y_1^v = 1) &\leq \left(1 - \left((1-p)^{h_1}(1-q)^{k_2}\right)^{T_4}\right) \\ &\leq 1 - e^{-T_4(ph_1 + qk_2)} \\ &\leq T_4(ph_1 + qk_2) \end{aligned}$$

So, since it holds $a_1 > 1 - a > 1 - b$, we get $\mathbf{E}[Y_1] \leq 2 n p h_1 T_4 = O(n^{a_1} \text{polylog } n)$. By repeating the same reasoning for the other 2 windows and by applying the Chernoff bound, the thesis follows. \square

For the sake of brevity, we define $k_1 = k_1^{(\tau_3)}$, $k_2 = k_2^{(\tau_3)}$, and $h = h_1^{(\tau_3)}$.

Let us consider a node $v \in V_1$ at the end of the first time window of Phase 4. For some constant $\gamma > 0$, it holds that

$$\begin{aligned} \mathbf{P}(X_1^v = 1) &\geq \left(1 - (1-p)^{k_1 T_4}\right) (1-p)^{h_1 T_4} (1-q)^{k_2 T_4} \\ &\geq \left(1 - \frac{\gamma}{\log^{1-\eta} n}\right) p k_1 T_4. \end{aligned}$$

Since $\frac{1}{n} \leq p \leq \frac{\log n}{n}$, by computing the expected value of the sum of all X_1^v 's and by applying the Chernoff bound, we get that the number of well-colored nodes in V_1 at the end of the first time window of Phase 4 is w.h.p.

$$k_1^{T_4} \geq d_4 \frac{n}{\log^2 n},$$

where $d_4 = d_4(c_4)$ is a positive constant that can be made arbitrarily large by increasing c_4 in $T_4 = c_4 \log n$.

We have thus shown that, after the first window, the number of well colored nodes inside each community is increased by a factor $d_4 \log n$. We can then repeat the same analysis for the second and the third windows (which are necessary when $p = o(\log n/n)$). Let us consider the sparsest case $p = 1/n$ (the other cases are easier). In this case, at the end of the third window, it can be easily verified that:

$$\begin{aligned} \mathbf{P}(X_1^v = 1) &\geq \left(1 - (1-p)^{\frac{n}{\log n} T_4}\right) (1-p)^{(n^{a_1} \text{polylog } n) T_4} (1-q)^{n T_4} \\ &\geq \left(1 - \frac{1}{n^\epsilon}\right) (1 - e^{-c_4}) \end{aligned}$$

The last bound can be thus made arbitrarily close to 1 by increasing the constant c_4 . Hence, w.h.p.

$$k_1^{(\tau_4)} \geq \alpha n$$

where constant α can be made arbitrarily close to 1 by suitably choosing the constant c_4 in $T_4 = c_4 \log n$.

As for the bad coloring, the thesis follows from Claim 8. \square

2.5 Phase 5: majority rule.

Theorem 4 states that, at the end of Phase 4, the Markovian Process w.h.p. reaches a state where a (large) constant fraction of the nodes (say, $3/4$) in both communities is well-colored while only $O(n^{a_1} \text{polylog } n)$ nodes are bad-colored. We now show that a further final phase, where nodes apply a simple majority rule, yields the good coloring, w.h.p.. *Every* node $v \in V$ applies the following coloring rule:

- For every $t \in [1, T_5 = c_5 \log n]$, every node v observes the colors of its neighbors at time t and, for every color z_i ($i = 1, 2$), v computes the number f_i^t of its neighbors colored with z_i .
- Then, node v gets color z_1 if $\sum_{t \in [1, \dots, \tau_5]} f_1^t \geq \sum_{t \in [1, \dots, \tau_5]} f_2^t$, otherwise v gets color z_2 .

Let us assume the Markovian Process starts Phase 5 from a state satisfying Eq. 16 (say with constant $\alpha = 3/4$).

Theorem 5 *A constant $c_5 > 0$ can be fixed so that, at time $\tau_5 = \tau_4 + c_5 \log n$, every node of each community is well-colored, w.h.p.*

Sketch of Proof. Let us consider a node $v \in V_1$ and, for every time step t of Phase 5, define the r.v. X_t^v counting the number of its z_1 -colored neighbors and the r.v. Y_t^v counting the number of its z_2 -colored neighbors in E_t . Then, define the two sums

$$X_v = \sum_{t \in [\tau_4+1, \dots, \tau_5]} X_t^v \quad \text{and} \quad Y_v = \sum_{t \in [\tau_4+1, \dots, \tau_5]} Y_t^v$$

Let us also define the subset

$$G^{\tau_4} = \{v \in V_1 \mid v \text{ is } z_1\text{-colored at time } \tau_4\}$$

Thanks to Condition 16 (with constant $\alpha = 3/4$), it holds that

$$|G^{\tau_4}| \geq \frac{3}{4} |V_1| = \frac{3}{8} n$$

From the above inequality, the expected values of r.v.s X_v and Y_v can be easily bound as follows

$$\begin{aligned} \mathbf{E}[X_v] &\geq \sum_{t \in [\tau_4+1, \dots, \tau_5]} \sum_{u \in G^{\tau_4}} \mathbf{P}((u, v) \in E_t) \geq \frac{3}{8} pn\tau_5, \quad \text{and} \\ \mathbf{E}[Y_v] &\leq \sum_{t \in [\tau_4+1, \dots, \tau_5]} \left(\sum_{u \notin G^{\tau_4}} \mathbf{P}((u, v) \in E_t) + \sum_{u \in V_2} \mathbf{P}((u, v) \in E_t) \right) \leq \frac{1}{7} pn\tau_5 \end{aligned}$$

Finally, observe that X_v and Y_v are sums of independent binary r.v.s (thanks to the $\mathcal{G}(n, p, q)$ model). Since $p \geq 1/n$ and $\tau_5 = \tau_4 + c_5 \log n$, we can thus choose a suitable constant $c_5 > 0$ and apply the Chernoff bound to get the thesis. \square

2.6 Overall completion time of the Protocol and its optimality

When p and q satisfy Cond. (1), we have shown that every phase has length $O(\log n)$: the Protocol has thus an overall completion time $O(\log n)$. In Section 3, we will show that for $p = o(1/n)$ the length of each phase must be stretched to $\Theta\left(\frac{\log n}{pn}\right)$.

As for the time optimality of our protocol we can state the following

Theorem 6 *If $p = O(1/n)$ and $q = O(p/n^b)$ for some constant $b > 0$, then any good-coloring protocol requires $\Omega(\log n)$ expected time.*

Sketch of Proof. if $p = O(1/n)$, starting from the initial random graph-snapshot, it is easy to show that there is non-negligible probability that some node will be isolated for $\tau(n)$ time steps where $\tau(n)$ is any increasing function such that $\tau = o\left(\frac{\log n}{pn}\right)$. It is clear that, in this time window, such isolated nodes cannot get a good color (w.h.p.)

□

3 More General Settings

In this section, we show some relevant generalizations that can be efficiently solved by simple adaptations of our protocol and/or its analysis.

- Removing the presence of two leaders. So far we have assumed that, in the initial state of the coloring process, there are exactly two source nodes, one in each community, which are colored with different colors. This assumption can be removed by introducing a preliminary phase in which a randomized source election is performed and by some further changes that are described below. In the first step, every node, by an independent random choice, becomes a *source* with probability $\frac{d \log n}{n}$ for a suitable constant $d > 0$. This clearly guarantees that, in every community, there are w.h.p. $\Theta(\log n)$ sources. Then, every source s_i randomly chooses a color $z_i \in [n^2]$. This implies that the minimal color z_1 in the first community and the minimal color z_2 in the second community are different w.h.p..

Let a and b be the number of sources chosen in V_1 and V_2 , respectively, and define $\ell = a + b$. We summarize the above arguments in the following

Fact 1 *Two positive constants $\eta_1 < \eta_2$ exist such that at the end of the first step w.h.p. it holds that $\eta_1 \log n \leq a, b \leq \eta_2 \log n$ and $z_1 \neq z_2$.*

The generic state of the modified Markovian Process is represented by the following set of random variables:

$$(a, b; k_1^1, \dots, k_a^1, h_1^1, \dots, h_b^1, k_1^2, \dots, k_b^2, h_1^2, \dots, h_a^2)$$

where k_j^i equals the number of nodes in V_i colored by the same (*good*) color as the j th source of V_i while h_j^i equals the number of nodes in V_i colored by the same (*bad*) color as the j th source of V_r with $r \neq i$. At every time step t , for any $v \in [n]$ we define the r.v. $N_j^v(t)$ as the the number of v -neighbors colored with color z_j at time t .

The first three phases of the Protocol are identical to the 2-source case since the impact of the presence of an $O(\log n)$ colors in each of the two communities remains negligible till the overall number of colored nodes in each community is $O(n/\log^4 n)$. By applying the same analysis of the 2-source case, at the end of Phase 3, we can thus show that the Markovian Process w.h.p. reaches a state having similar properties to those stated in Theorem 3. We remind that p and q belong to the ranges in Cond. (1).

Theorem 7 *We can choose a suitable $\tau_3 = \tau_2 + (c_3 + o(1)) \log n$ so that, at the end of Phase 3, the Markovian Process w.h.p. reaches a state in which for $\ell = 1, 2$ it holds*

$$\forall j \in [a] \quad \frac{n}{\log^4 n} \leq k_j^\ell \leq \frac{n}{\log^{4-\eta} n} ; \quad \forall i \in [b] \quad h_i^\ell = \sqrt{n} \text{ polylog } n$$

where η is a constant that can be made arbitrarily small.

We need to stop at a “saturation level” $O(n/\log^{4-\eta} n)$ for *every* good color, since we want to guarantee (w.h.p.) that the minimal color infects at least $n/\text{polylog } n$ nodes. Then, as in the 2-source case, the protocol starts a controlled saturation phase (i.e. Phase 4) that consists of (at most) 4 consecutive time windows in which every node applies the same following *minimal-color* rule:

For $t = 1$ to $T_4 = c_4 \log n$ time steps, v observes the colors of its neighbors and gets the *minimal* color \hat{z} among all the observed colors.

Thanks to the above rule, the size of the nodes colored by the minimal good-color increases by a logarithmic factor at the end of each of the four windows. This fact can be proved by using the same arguments of the proof of Theorem 4.

It thus follows that, at the end of Phase 4, the number of nodes colored with the good minimal color is at least a constant (say 3/4) fraction of all the nodes of the community. Then, as in the 2-source case, every node can apply the majority rule in order to get the right color w.h.p.

- The case p -unknown. Our protocol relies on the fact that nodes know the parameter $p = \frac{d}{n}$: the length of the protocol’s phases are functions of p . So an interesting issue is to consider the scenario where nodes do not know the parameter p (i.e. the expected degree). Thanks to edge independence, the dynamic random-graph process can be seen by every node as an independent sequence of random samples. Indeed, at every time step t , every node can store the number $|N^v(t)|$ of its neighbors and it knows that this number has been selected by $n - 1$ independent experiments according to the same Bernoulli distribution with success probability $p = \frac{d}{n}$. The goal is thus to use such samples in order to get a good approximation of p . If $p \geq \frac{1}{n}$, by using a standard statistical argument, every node w.h.p. will get the value of p up to some negligible factor in $O(\log n)$ time. Let’s see this task more formally.

For $c \log n$ time steps (where c is a constant that will be fixed later), every node stores the values $|N^v(1)|, |N^v(2)|, \dots, |N^v(c \log n)|$; then it computes $S = |N^v(1)| + \dots + |N^v(c \log n)|$. Since S is the sum of $c \log n \cdot (n - 1)$ Bernoulli random variables of parameter $\frac{d}{n}$, we get a binomial distribution with mean $\mathbf{E}[S] = dc \log n (1 - \frac{1}{n})$. Then, every node uses the estimator $D(S) = \frac{S}{c \log n (1 - \frac{1}{n})}$ to guess d . We can use the Chernoff bound in order to determine a confidence interval for $D(S)$, as follows

$$\begin{aligned} \mathbf{P}(d \notin [D(S) - \delta, D(S) + \delta]) &= \mathbf{P}\left(S \notin [\mathbf{E}[S] - \delta c \log n \frac{n-1}{n}, \mathbf{E}[S] + \delta c \log n \frac{n-1}{n}]\right) = \\ \mathbf{P}\left(S < \mathbf{E}[S] \left(1 - \frac{\delta}{d}\right)\right) &+ \mathbf{P}\left(S > \mathbf{E}[S] \left(1 + \frac{\delta}{d}\right)\right) < e^{-\frac{\delta^2}{2d^2} \mathbf{E}[S]} + e^{-\frac{\delta^2}{3d^2} \mathbf{E}[S]} < 4 \left(\frac{1}{n^c}\right)^{\frac{\delta^2}{3d}} \end{aligned}$$

It thus follows that, for any $d \geq 1$, we can choose $\delta = \sqrt{d}$ and c sufficiently large in order to get a good confidence interval for all nodes of the network. This obtained approximation suffices to perform an analysis of the protocol which is equivalent to that of the case p -known.

- Edge Markovian Evolving Graphs. Let us consider an Edge-MEG $\mathcal{G}(n, p_\uparrow, p_\downarrow, q_\uparrow, q_\downarrow, E_0)$ defined in the introduction and assume that $q_\uparrow \leq p_\uparrow/n$. If $0 < p_\uparrow, p_\downarrow, q_\uparrow, q_\downarrow < 1$, it is easy to see [6] that the (unique) stationary distribution of the two corresponding 2-state edge-Markov chains (inside and outside the communities, respectively) are

$$\pi^{\text{in}} = \left(\frac{p_\downarrow}{p_\uparrow + p_\downarrow}, \frac{p_\uparrow}{p_\uparrow + p_\downarrow} \right) \quad \text{and} \quad \pi^{\text{out}} = \left(\frac{q_\downarrow}{q_\uparrow + q_\downarrow}, \frac{q_\uparrow}{q_\uparrow + q_\downarrow} \right)$$

It thus follows that the dynamic graph, starting from any E_0 , converges to the (2-communities) Erdős-Rényi random graph with edge-probability functions

$$\tilde{p} = \frac{p_{\uparrow}}{p_{\uparrow} + p_{\downarrow}} \quad (\text{Inside Communities}) \quad \text{and} \quad \tilde{q} = \frac{q_{\uparrow}}{q_{\uparrow} + q_{\downarrow}} \quad (\text{Outside Communities})$$

The mixing time M^{in} and M^{out} of the two edge Markov chain are bounded by [6] $M^{in} = O\left(\frac{1}{p_{\uparrow} + p_{\downarrow}}\right)$, $M^{out} = O\left(\frac{1}{q_{\uparrow} + q_{\downarrow}}\right)$. Let us observe that there is a *Markovian dependence* between graphs of consecutive time steps. If we observe any event at time t related to E_t (such as the number of well-colored nodes) then E_{t+1} is not anymore random with the stationary distribution.

It thus follows that we need to change the way the protocol works over the dynamic random graph. Let $M = \max\{M^{in}, M^{out}, \log n\}$; then by definition of mixing time, starting from *any* edge subset E_t at time t , at time $t + \Delta$ with some $\Delta = \Theta(M)$, if $u, v \in V_1$ or $u, v \in V_2$ then edge (u, v) exists with probability $\tilde{p} \pm \frac{1}{n^2}$, otherwise it exists with probability $\tilde{q} \pm \frac{1}{n^2}$. In other words, whatever the state of the coloring process is at time t , after a time window proportional to the mixing time, the dynamic graph is random with a distribution which is very close to the stationary one. We can thus modify our protocol for the *dynamic Erdős-Rényi graph* model $\mathcal{G}(n, p, q)$ in order to “wait for mixing”. Between any two consecutive steps of the original protocol there is a *quiescent* time-window of length $\Theta(M)$ where every node simply does nothing. Then, the analysis of the protocol over $\mathcal{G}(n, p_{\uparrow}, p_{\downarrow}, q_{\uparrow}, q_{\downarrow}, E_0)$ is similar to that in Section 2 working for the *dynamic Erdős-Rényi graph* $\mathcal{G}(n, \tilde{p}, \tilde{q})$. We can thus state that, under the condition $q_{\uparrow} \leq O(p_{\uparrow}/n^b)$ for some constant $b > 0$, this version of our protocol w.h.p. performs a good-coloring in time $O\left(M \cdot \max\left\{\log n, \frac{\log n}{pn}\right\}\right)$. We finally observe that, for the “realistic” case $p_{\downarrow}, q_{\downarrow} = \Theta(1)$ (see the discussion in the Introduction), the mixing-time bound M turns out to be $O(\log n)$: we thus get only a logarithmic slowdown-factor w.r.t. the good-coloring in the *dynamic Erdős-Rényi graph* $\mathcal{G}(n, \tilde{p}, \tilde{q})$.

- **More Communities.** The presence of a constant number $r = \Theta(1)$ of unknown equally-sized communities can be managed with a similar method to that described above for removing the presence of leaders. Indeed, the major issue to cope with is the presence of a constant number of different color spreadings in each community and the protocol must select the right one in every community. However, if r is a constant and the number of nodes in each community is some constant fraction of n , then the impact of the presence of $O(\log n)$ colors in each of the r communities remains negligible till the overall number of colored nodes in each community is $O(n/\log^4 n)$. As in the previous paragraph, by first applying the minimal-color rule and then the majority one, the modified protocol returns a good-coloring w.h.p.

- **Sparse Graphs.** When $p = o\left(\frac{1}{n}\right)$ and $\frac{q}{p} = O\left(\frac{1}{n^b}\right)$ (for some constant $b > 0$), the snapshots of the dynamic graph are very sparse. So, every node must wait at least $\Theta\left(\frac{1}{pn}\right)$ time step (in average) in order to meet some other node. This implies that the coloring protocol will be slower. We can reduce this case to the case $p = 1/n$ by considering the *time-union* random graph obtained from $\mathcal{G}(n, p, q)$ according to the following

Definition 8 Let Δ be any positive integer and consider any sequence of graphs $G(V, E_1), \dots, G(V, E_{\Delta})$. Then, we define the Δ -OR-graph

$$G_{\Delta}^{\Delta} = (V, E^{\Delta}) \quad \text{where} \quad E^{\Delta} = \left\{e \in \binom{V}{2} \mid \exists t^{\star} \in (1, \Delta] : e \in E_{t^{\star}}\right\}$$

It is easy to prove the following

Lemma 9 *Let $p < \frac{1}{n}$, then the $\frac{1}{pn}$ -OR-graph of any finite sequence of graphs selected according to the $\mathcal{G}(n, p, q)$ model is a $\mathcal{G}(n, \tilde{p}, \tilde{q})$ with $\tilde{p} = \Theta\left(\frac{1}{n}\right)$ and $\tilde{q} = O\left(\frac{p}{n^b}\right)$.*

The modified protocol just works as it would work over $\mathcal{G}(n, \tilde{p}, \tilde{q})$ with $\tilde{p} = \Theta\left(\frac{1}{n}\right)$ and $\tilde{q} = O\left(\frac{p}{n^b}\right)$: in every phase, every node applies the phase's coloring rule (only) every $\Delta = \frac{1}{pn}$ time steps on the $\frac{1}{pn}$ -OR-graph. The modified protocol thus requires $\Theta(\Delta \log n) = \Theta\left(\frac{\log n}{pn}\right)$ time.

- **Dense Graphs.** When p becomes larger than $\log n/n$ and $q = \left(\frac{p}{n^b}\right)$, the coloring problem becomes an easier task since standard probability arguments easily show that the (good) coloring process is faster and the related random variables (i.e. number of new colored nodes at every time step) have much smaller variance. This implies that the protocol can be simplified: for instance, the source-coloring phase (i.e. Phase 1) can be skipped while the length of other phases can be reduced significantly as a function of p . However, we again emphasize that *dense* dynamic random graphs are not a good model for the scenario we are inspired from: ICMNs are opportunistic networks having sparse and disconnected topology.

4 Conclusions

This paper introduces a framework that allows an analytical study of the distributed community-detection problem in dynamic graphs. Then, it shows an efficient algorithmic solution in two classes of such graphs that model some features of opportunistic networks such as ICMNs. We believe that the problem deserves to be studied in other classes of dynamic graphs that may capture further relevant features of social opportunistic networks.

Acknowledgements. We thank Stefano Leucci for its help in getting an efficient protocol simulation over large random graphs.

References

- [1] C. Avin and M. Koucky and Z. Lotker. How to explore a fast-changing world. In *Proc. of 35th ICALP'08*, LNCS, 5125, 121–132, 2008.
- [2] H. Baumann, P. Crescenzi, and P. Fraigniaud. Parsimonious flooding in dynamic graphs. In *Proc. of the 28th ACM PODC '09*, 2009.
- [3] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, J. Scott, and R. Gass. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Trans. Mob. Comp.*, 6(6), 2007.
- [4] A. Chaintreau, A. Mtibaa, L. Massoulié, and C. Diot. Diameter of opportunistic mobile networks. In *Proc. of ACM Sigcomm CoNext'07*, 2007.
- [5] A. Clementi, C. Macci, A. Monti, F. Pasquale, R. Silvestri “Flooding Time of Edge-Markovian Evolving Graphs”. *SIAM J. Discrete Math.* 24(4): 1694-1712 (2010) (Ext. Abs. in *ACM PODC* 2008).

- [6] A. Clementi, A. Monti, F. Pasquale, and R. Silvestri. “Information spreading in stationary markovian evolving graphs. *IEEE Trans. Parallel Distrib. Syst.*, 22(9): 1425–1432, 2011 (Ext. Abs. in *IEEE IPDPS* 2009).
- [7] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. ArXiv TR: arXiv:cond-mat-0505245 [cond-mat.dis-nn], DOI: 10.1088/1742-5468/2005/09/P090082005.
- [8] N. Eagle and A. Pentland. Reality mining: Sensing complex social systems. *Journal of Personal and Ubiquitous Computing*, 10(4):255–268, 2005.
- [9] M. Handcock, A. Raftery and J. Tantrum. Model-based clustering for social networks *J. R. Statist. Soc. A*, 170, 301–354, 2007
- [10] P. Hui, E. Yoneki, S. Yan Chan, and J. Crowcroft. Distributed community detection in delay tolerant networks. In *Proc. of 2nd ACM/IEEE MobiArch ’07*, 2007.
- [11] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnovic. Power Law and Exponential Decay of Inter Contact Times Between Mobile Devices. In *Proc. 13th ACM MOBICOM*, 183–194, 2007.
- [12] F. Martelli, M. Renda, G. Resta, and P. Santi. A Measurement-based Study of Beaconing Performance in IEEE 802.11p Vehicular Networks. *Proc. of IEEE INFOCOM’12*, 2012.
- [13] T. Spyropoulos, A. Jindal, and K. Psounis. An analytical study of fundamental mobility properties for encounter based protocols. *Int. J. Auton. Adapt. Commun. Syst.*, 1(1), 4–40, 2008.
- [14] P.-U. Tournoux, J. Leguay, F. Benbadis, V. Conan, M. D. de Amorim, and J. Whitbeck. The accordion phenomenon: Analysis, characterization, and impact on DTN routing. in *Proc. IEEE INFOCOM’09*, 2009.
- [15] M. Vojnovic and A. Proutiere, A. Hop limited flooding over dynamic networks. *Proc. of IEEE INFOCOM’11*, 685-693, 2011.
- [16] Whitbeck, J.; Conan, V.; de Amorim, M.D. Performance of Opportunistic Epidemic Routing on Edge-Markovian Dynamic Graphs. *Communications, IEEE Transactions on* , vol.59, no.5, pp.1259-1263, 2011.
- [17] Yoneki E., Hui P., and Crowcroft J. Wireless Epidemic Spread in Dynamic Human Networks. In *Bio-Inspired Computing and Communication*, LNCS, 5151, 116-132 2008.

A Experimental Results

We run our protocol over sequences of independent random graphs according to the $\mathcal{G}(n, p, q)$ model. The protocol has been suitably simplified and tuned in order to optimize the real performance. In particular, the implemented protocol consists of 5 Phases: Phase 1 (Source-Coloring), Phase 2-3 (Fast-Coloring I-II), Phase 4 (Min-Coloring), and Phase 5 (Majority-Rule). The rules of each phase is the same of the corresponding phase analyzed in Section 2. Moreover, the length of every phase is fixed to $c \log n$. As shown in the next tables, parameter c is always very small and it depends on the parameter q . The parameter c has been heuristically chosen as the minimal one yielding the good coloring in more than 98% of the trials. We consider instances of increasing size n and for each size, we tested 100 random graphs. In the first experiment class (see Table 1), we consider homogeneous sparse graphs with the following setting: $p = \frac{5}{n}$ and 3 values of q ranging from $1/n^2$ to $1/n^{3/2}$.

Table 1: Tab. 1. Experimental results for the homogeneous case. For every value of n , the rows indicates the percentage of good-coloring for three choices of q and the “minimal” setting for c (the total number of Protocol’ steps is inside brackets).

n	$q = n^{-\frac{3}{2}}, c = 0.9$	$q = n^{-\frac{5}{3}}, c = 0.6$	$q = n^{-2}, c = 0.5$
20000	99 (66)	100 (46)	100 (36)
40000	99 (71)	100 (46)	100 (41)
80000	100 (76)	100 (51)	100 (41)
160000	100 (81)	100 (51)	100 (46)
320000	100 (86)	100 (56)	99 (46)
640000	100 (91)	100 (61)	100 (51)
1280000	100 (91)	100 (61)	100 (51)
2560000	100 (96)	100 (66)	100 (56)

The second class of experiments concerns non-homogeneous random graphs. For each pair of nodes $e = (u, v)$ in the same community, the probability p_e is randomly fixed in a range $[d_1/n, d_2/n]$ before starting the graph-sequence generation. Then, at every time step $t \geq 0$, the graph-snapshot $G(V, E_t)$ is generated by selecting every edge $e = (u, v)$ according to its birth-probability p_e (the edges between the two communities are generated with parameter q). In Table 2, the experimental results are shown for the case $d_1 = 1$ and $d_2 = 9$ in order to generate sparse topologies inside the communities, while in Table 3, the results concern the more dense case where $d_1 = 0$ and $d_2 = \log n$. The protocol’s implementation is the same of the homogeneous case above.

The experiments globally show that the tuning of parameter c mainly depends on the value of q even though it can be fixed to small values in all studied cases. Moreover, the presence of non-homogeneous edge-probability function seems to slightly “help” the efficiency of the protocol. Intuitively speaking, we believe this is due to the presence of fully-random irregularities in the graph topology that helps the protocol to *break* the symmetry of the initial configuration.

Table 2: Tab. 2. Experimental results for the non-homogeneous sparse case with $d_1 = 1$ and $d_2 = 9$.

n	$q = 1/n^{3/2}, c = 1$	$q = 1/n^{5/3}, c = 0.4$	$q = 1/n^2, c = 0.4$
20000	100 (46)	100 (46)	100 (36)
40000	98 (71)	99 (46)	100 (41)
80000	100 (76)	100 (51)	100 (41)
160000	100 (81)	100 (51)	100 (46)
320000	100 (86)	100 (56)	100 (46)
640000	100 (91)	100 (61)	100 (51)
1280000	100 (91)	100 (61)	100 (51)

Table 3: Tab. 3. Experimental results for the non-homogeneous case with $d_1 = 0$ and $d_2 = \log n$.

n	$q = n^{-\frac{3}{2}}, c = 1$	$q = n^{-\frac{5}{3}}, c = 0.4$	$q = n^{-2}, c = 0.4$
20000	99 (76)	100 (31)	100 (31)
40000	99 (81)	100 (31)	100 (31)
80000	98 (86)	100 (31)	100 (31)
160000	100 (91)	100 (36)	100 (36)
320000	100 (96)	100 (36)	100 (36)
640000	100 (101)	100 (41)	100 (41)
1280000	100 (106)	100 (41)	100 (41)

B Useful Tools

Lemma 10 *If $x = o(1)$ and $xy = o(1)$ then*

$$\begin{aligned} (1-x)^y &\geq 1 - xy(1+2x) \\ (1-x)^y &\leq 1 - xy(1-xy) \end{aligned}$$

We will often use the Chernoff's bounds

Lemma 11 (Chernoff's Bound.) *Let be $X = \sum_{i=1}^n X_i$ where X_1, \dots, X_n are independent Bernoulli random variables and let be $0 < \delta < 1$. If $0 < \mu_1 \leq \mathbf{E}[X]$ and $\mu_2 \geq \mathbf{E}[X]$, then it holds that*

$$\mathbf{P}\{X \leq (1-\delta)\mu_1\} \leq e^{-\frac{\delta^2}{2}\mu_1}. \quad (17)$$

$$\mathbf{P}(X \geq (1+\delta)\mu_2) \leq e^{-\frac{\delta^2}{3}\mu_1}. \quad (18)$$

Lemma 12 *Let φ be any poly-logarithm and $E_0, E_1, \dots, E_\varphi$ be events that hold w.h.p., then $E_0 \cap E_1 \cap \dots \cap E_\varphi$ holds w.h.p.*